

Article

Not peer-reviewed version

# Optical Remote Sensing Ship Classification and Recognition based on Improved YOLOv5

[Jun Jian](#)<sup>\*</sup>, [Long Liu](#), Yingxiang Zhang, Ke Xu, [Jiaxuan Yang](#)

Posted Date: 4 July 2023

doi: 10.20944/preprints202307.0150.v1

Keywords: optical remote sensing images; convolutional block attention module; cross-layer connection channel; lightweight GSConv; Wise-IoU loss function; median + bilateral filter; object detection



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# Optical Remote Sensing Ship Classification and Recognition Based on Improved YOLOv5

Jun Jian <sup>1,\*</sup>, Long Liu <sup>1</sup>, Yingxiang Zhang <sup>1</sup>, Ke Xu <sup>1</sup> and Jiaxuan Yang <sup>1</sup>

<sup>1</sup> Navigation College, Dalian Maritime University, Dalian 116026, China; jianjun@dlmu.edu.cn (J.J.); ll1120211048@dlmu.edu.cn (L.L.); zyxmz@dlmu.edu.cn (Y.Z.); xuke0225@dlmu.edu.cn (K.X.); yangjiaxuan@dlmu.edu.cn (J.Y.)

\* Correspondence: jianjun@dlmu.edu.cn (J.J.)

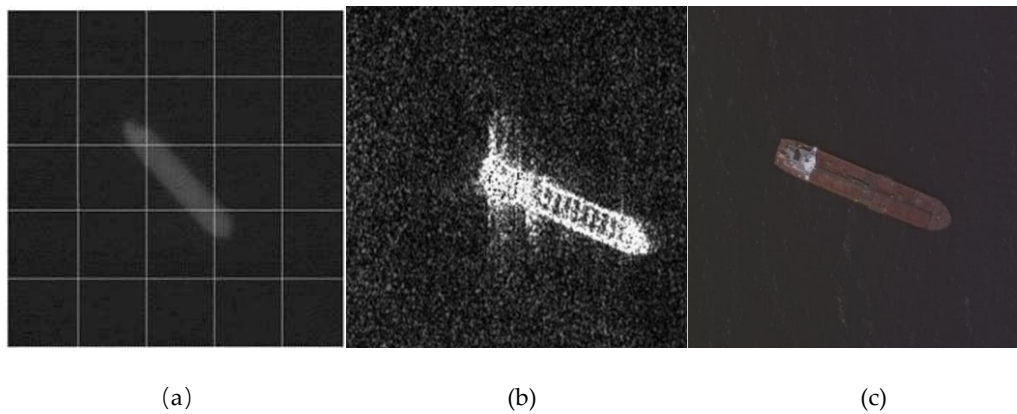
**Abstract:** Due to the special characteristics of the shooting distance and angle of remote sensing satellites, the pixel area ratio of ship targets is small and the feature expression is insufficient, which leads to unsatisfactory ship detection performance and even situations such as missed detection and false detection. In this study, we propose an improved-YOLOv5 algorithm. The improvement strategies mainly include: (1) Add the Convolutional Block Attention Module (CBAM) into the Backbone to enhance the extraction of target-adaptive optimal features; (2) Introduce cross-layer connection channel and lightweight GSConv structure into the Neck to achieve higher-level multi-scale feature fusion and reduce the number of model parameters; (3) The Wise-IoU loss function is used to calculate the localization loss in the Output, and assign reasonable gradient gains to cope with differences in image quality. In addition, during the preprocessing stage of experimental data, a median and bilateral filter method is used for noise reduction to reduce interference from ripples and waves and highlight the information of ship features. The experimental results show that Improved-YOLOv5 has a significant improvement in recognition accuracy compared to various mainstream target detection algorithms; Compared to the original YOLOv5s, the mean Average Precision (mAP) has improved by 3.2% and the Frames Per Second (FPN) has accelerated by 8.7%.

**Keywords:** optical remote sensing images; convolutional block attention module; cross-layer connection channel; lightweight GSConv; Wise-IoU loss function; median + bilateral filter; object detection

## 1. Introduction

In recent years, with the rapid development of remote sensing satellite technology, people have been brought into an era of comprehensive, multi-angle, and three-dimensional observation of the Earth. Remote sensing satellites have become an important means of observing ship targets on the ocean surface due to their unique advantages, while also providing a large number of high-resolution remote sensing images.

Ship target detection is mainly based on three types of images. One type is infrared remote sensing images, another is synthetic aperture radar (SAR), and the third is optical remote sensing images. Infrared remote sensing images have strong environmental adaptability and long detection range, but suffer from low image contrast and poor detail resolution [1–3]. SAR images have the advantages of all-weather and robustness, and can detect ship targets in the presence of clutter or noise interference. However, the images lack detail and color information, making it difficult to identify ship types [4]. Optical remote sensing images typically use electromagnetic wave imaging principles and can provide more detailed information about ships, which is beneficial for ship target identification and classification [5]. It can effectively complement the above two types of images. The imaging effects of these three remote sensing images are shown in **Error! Reference source not found.** Therefore, ship target detection based on optical remote sensing images has become an important method for monitoring ships.



**Figure 1.** The imaging effects of three remote sensing images: (a) infrared remote sensing image; (b) SAR image; (c) optical remote sensing image.

Ships are the primary carriers for maritime cargo transportation and critical targets in military activities. Therefore, ship recognition and classification based on optical remote sensing imagery plays an increasingly important role in various maritime affairs [4]. Civilians can use it to monitor ship traffic conditions, prevent congestion, aid in emergency search and rescue operations, command vessels in special sea areas and nearby ports, and combat illegal activities like pollution, smuggling, and human trafficking effectively. Militarily, it is helpful to grasp the distribution of enemy ships and make strategic deployments promptly. This provides a reference for winning sea battles.

With the application and expansion of deep learning in the field of computer vision, deep learning-based target detection methods have gradually replaced traditional handcrafted feature design and become a research hotspot in various industries in society. The early target detection algorithms relied heavily on manual feature extraction, traversed the image through sliding windows and finally determined the target class by classifiers, which has limitations such as high computational effort, slow speed, high error rate and strong subjectivity [6]; In contrast, object detection algorithms based on deep learning offer high operational efficiency, fast processing speed, and high detection accuracy, making them suitable for real-time detection. The use of computer vision and convolutional neural network (CNN) has become the dominant algorithm for ship target detection. Deep learning target detection algorithms can be divided into ONE-STAGE and TWO-STAGE [7]. TWO-STAGE algorithms generate first region proposal and then subject to CNN for target classification and position regression. Typical models include R-CNN [8], Fast R-CNN [9] and Faster R-CNN [10], etc. Although this method achieves high detection accuracy, it is relatively slow. ONE-STAGE algorithms directly classify and predict the position of targets without generating region boxes, represented by SSD [11] and the YOLO series [12], etc. Compared with the TWO-STAGE algorithms, its detection speed is faster, but its detection accuracy is reduced.

Currently, scholars have conducted research on ship target detection based on these two types of algorithms. Zhang P. P. et al. [13] added the residual convolution module in Faster-RCNN to improve feature representation ability; Meanwhile, the K-means method was introduced to cluster the size and aspect ratio of ship targets. Wen G. Q. et al. [14] proposed a multi-scale single-shot detector (MS-SSD) by introducing more high-level context and more appropriate supervision to improve the detection effect of small ship targets and enhance the model's robustness to scale variance. Chen L. Q. et al. [15] designed a novel and lightweight Dilated Attention Module (DAM) on the YOLOv3 benchmark framework to extract discriminative features for ship targets. It was aiming to detect ships with different scales in different backgrounds, while at a real-time speed. Huang Z. X. et al. [16] proposed an improved YOLOv4 ship target detection algorithm that introduces the Receptive Field Module (RFB) instead of Spatial Pyramid Pooling (SPP) to enlarge the receptive field and improve the detection of small targets. Zhou J. C. et al. [17] improved the YOLOv5s algorithm, which used the K-means clustering algorithm to re-cluster the target box, while added a

maximum pooling layer in the space pyramid to improve multiple receptive fields fusion, and finally used the C-IoU loss function to increase the restriction mechanism for the aspect ratio.

Compared with existing natural images, remote sensing images face problems such as diverse scales, a majority of small targets and closely arranged partial targets due to the differences in shooting distances and angles, so the obtained target feature information is relatively limited [4]. Moreover, remote sensing images are susceptible to interference from environmental factors like wave noise, resulting in the overall poor detection performance of current algorithms and even cases of missed and false detection [18].

In response to the above, this paper proposes an improved optical remote sensing ship target detection algorithm based on YOLOv5 (Improved-YOLOv5). The main work is as follows:

- Adding the Convolutional Block Attention Module (CBAM) [19] in the backbone network to focus on regions of interest, suppress useless information and improve the feature extraction capability;
- Inspired by the Weighted Bi-directional Feature Pyramid Network (BiFPN) [20], adding additional cross-layer connection channels in the Neck to enhance multi-scale feature fusion. Moreover, the lightweight GSConv structure [21] is introduced to replace conventional Conv, reducing the model parameters and accelerating convergence speed.
- The Wise-IoU loss function [22] is employed as the bounding box loss function at the Output to reduce the competitiveness of high-quality anchor boxes and mask the harmful gradients of low-quality examples.
- During the preprocessing stage of experimental data, a median filter and bilateral filter are used to reduce noise, such as water ripples and waves, and to highlight the ship feature information.

The above improvement strategies can effectively improve the problem of low accuracy of fine-grained classification recognition of multi-scale and small targets in complex scenes.

## 2. YOLOv5 Target Detection Algorithm

YOLOv5 as a classic model of ONE-STAGE algorithms, has a relatively fast detection speed. Based on differences in network depth and width, there are four versions of YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x, among which YOLOv5s is the smallest network and has the fastest running speed [23]. In this study, YOLOv5s is chosen as the baseline model to carry out improvement work, so as to meet the requirements of real-time ship target detection.

### 2.1. Network Structure

The YOLOv5s consists of four main parts: Input, Backbone, Neck, and Head [23]. **Error! Reference source not found.** illustrates the network structure diagram.

The input component mainly performs preprocessing data operations, including Mosaic data enhancement, adaptive anchor frame calculation and adaptive image scaling [24]. These techniques improve the training speed and network accuracy of the model.

The Backbone component adopts CSPDarknet53 as the backbone network [25], which mainly includes Conv (Conv2d+BN+SiLU) structure, C3 structure, and SSPF module. The network prevents overfitting and accelerates model convergence through Batch Normalization (BN); The C3 structure integrates gradient changes into the feature map to reduce computation while maintaining accuracy; The SSPF is an enhanced version of Spatial Pyramid Pooling (SPP), which further improves the running speed while preserving the original function.

The Neck component adopts the FPN+PAN structure, with FPN [26] passing semantic information from top to bottom and PAN [27] transmitting low-level semantic and positioning information from bottom to top, thereby enhancing semantic expression and positioning capability of multiple scales.

The Head component predicts target features, generates bounding boxes, and identifies the target category.

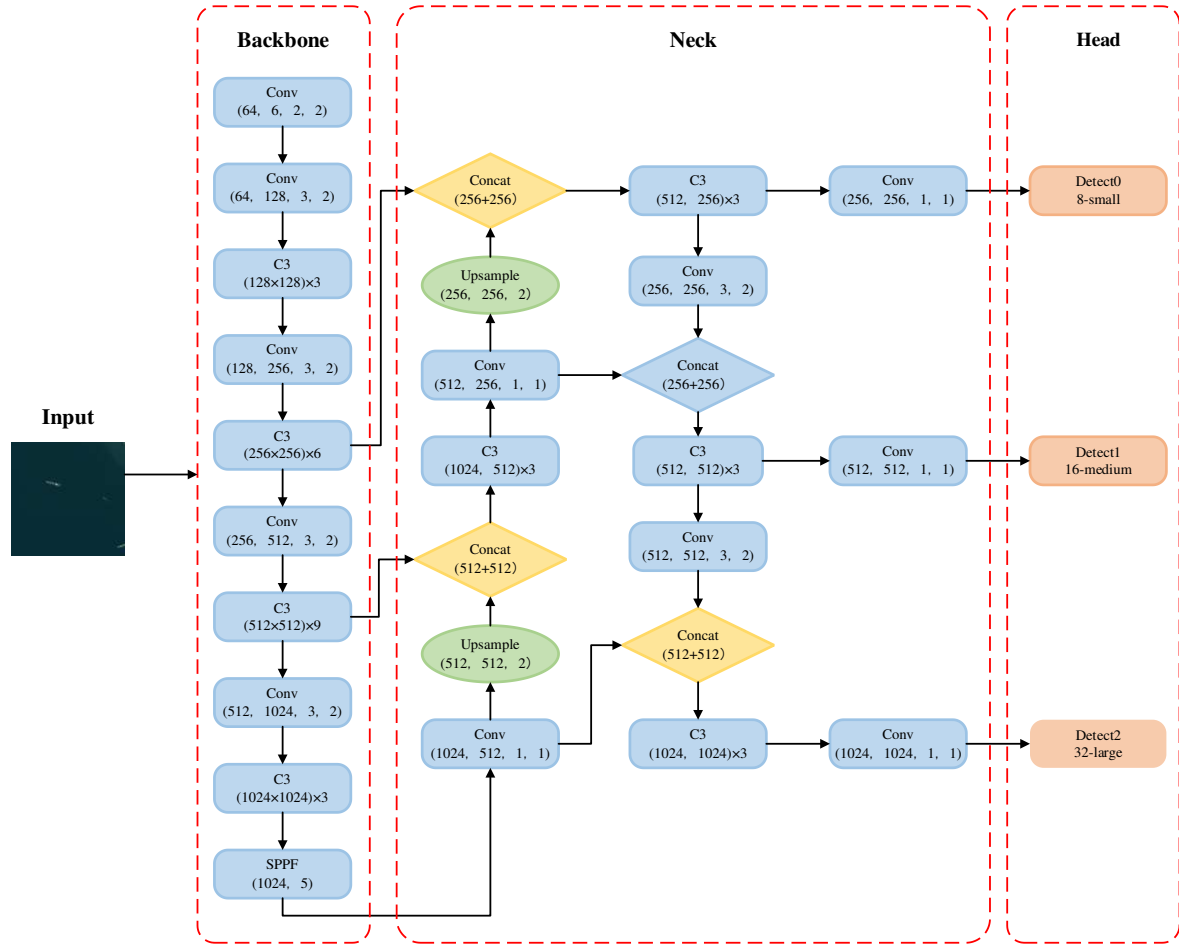


Figure 2. YOLOv5 network structure.

## 2.2. Loss Function

The YOLOv5 loss can be divided into three main components: classification loss, objectness loss, and localization loss [28]. The overall loss is shown in Equation (1).

$$L = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc} \quad (1)$$

where  $\lambda$  is the balance coefficient with values of 0.5, 1 and 0.05 respectively.

YOLOv5 used Binary Cross Entropy Loss function (BCE Loss) to calculate classification loss and objectness loss. BCE loss is defined as:

$$L_{BCE} = -y \log p - (1 - y) \log (1 - p) = \begin{cases} -\log p, & y = 1 \\ -\log (1 - p), & y = 0 \end{cases} \quad (2)$$

where  $y$  represents the label of the input sample (positive sample is 1, negative sample is 0);  $p$  represents the probability of the model predicting the input sample as a positive sample. Assuming

that  $p_t = \begin{cases} p, & y = 1 \\ 1 - p, & y = 0 \end{cases}$ , the definition of BCE Loss Equation can be simplified as:

$$L_{BCE} = -\log p_t \quad (3)$$

The localization loss employs the C-IoU loss function, which takes into account geometric relationships such as overlap area, centroid distance and aspect ratio, and also effectively addresses the divergence problem during subsequent training. Its definition is mainly calculated from the Intersection over Union (IoU) [29], as shown in Equation (4).

$$L_{C-IoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (4)$$

where:

$$\alpha = \frac{v}{(1 - IoU) + v}$$

$$v = \frac{4}{\pi^2} \left( \arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2$$

where  $\rho^2(b, b^{gt})$  is the Euclidean distance between the center point of the prediction box and ground truth box, also known as  $b$ ;  $c$  is the diagonal length of the minimum bounding box between the prediction and ground truth box;  $\alpha$  is the weight coefficient;  $v$  is used to measure the consistency of the aspect ratio;  $gt$  is the ground truth box;  $w$  and  $h$  are the width and height of the prediction box respectively.

### 3. Improved-YOLOv5 target detection algorithm

The paper proposes an Improved-YOLOv5 algorithm for ship target detection in optical remote sensing imagery, with the overall network architecture as shown in **Error! Reference source not found..**

During the improvement process of the backbone network, the CBAM module is added, which combines channel and spatial attention modules to make the feature extraction focus more on the target area and suppress useless information. In the neck structure, the BiFPN is used to introduce contextual [30] and weight information to balance features of different scales of ideas. Additional cross-layer connection channels are added to generate a larger receptive field and richer semantic information, achieving higher-level multi-scale feature fusion. Meanwhile, all Conv modules were replaced by GSConv modules to reduce the parameters and computation brought about by the feature pyramid structure upgrades [5]. Finally, the Wise-IoU loss function is used to calculate the bounding box loss in the Output. By re-evaluating the quality of anchor boxes, a wise gradient gain allocation strategy is provided. This strategy can effectively reduce the competitiveness of high-quality anchor boxes while also reducing harmful gradients generated by low-quality examples [22].



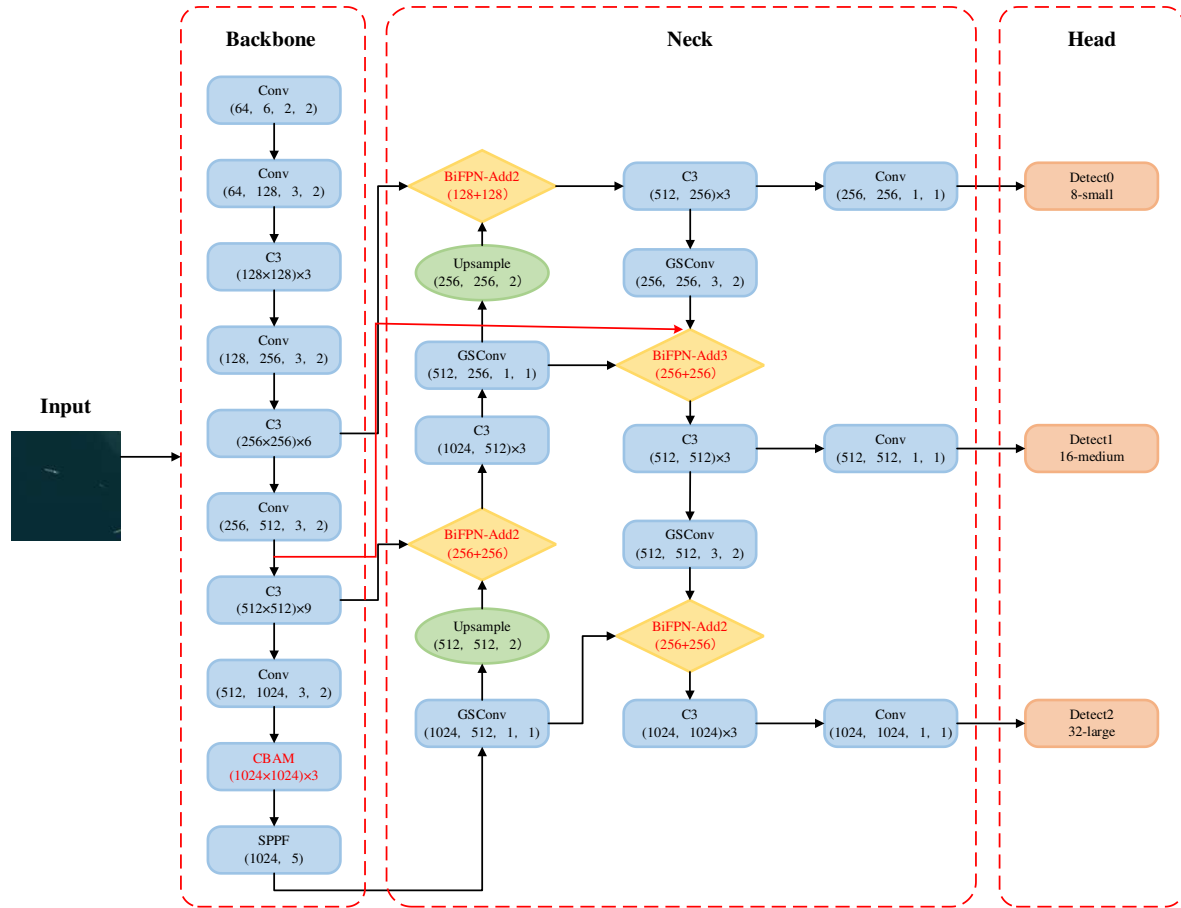


Figure 3. Improved-YOLOv5 network structure.

### 3.1. CBAM Attention Module

In complex scenarios, the majority of ship detection targets in remote sensing imagery occupy a relatively small proportion of the entire image, and the feature information extracted is small and inconspicuous, so the CBAM is added to the backbone network to obtain information of interest, suppress useless information and enhance the ability to extract feature information for small targets. Its working principle is shown in **Error! Reference source not found.**

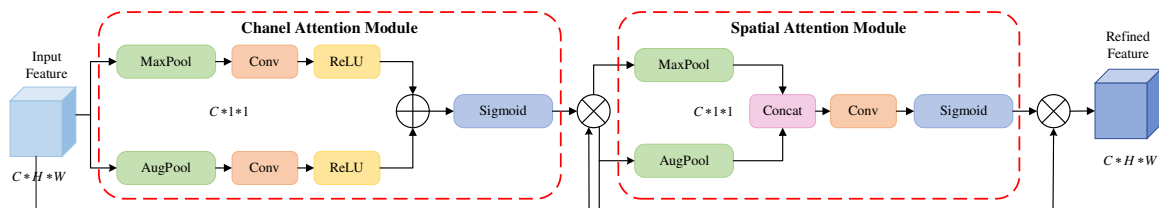


Figure 4. The structure of Convolutional Block Attention Module.

The CBAM consists of two main sub-modules, Channel Attention Module (CAM) and Spatial Attention Module (SAM) [31]. How it works: Firstly, CAM changes the feature map from  $C \times H \times W$  to  $C \times 1 \times 1$  through maximum pooling and average pooling, then after the Conv module and the ReLU activation function, the two activated results are summed element by element, then the output of the CAM is obtained through the Sigmoid activation function, and finally multiplied with the original map, which becomes the  $C \times H \times W$  size again. Secondly, SAM takes the output of the CAM as input, and also passes through the MaxPool and AugPool layers to obtain two  $C \times 1 \times 1$  feature maps, which are transformed into a 1-channel feature map through concatenating and  $7 \times 7$  convolution, and then passes through the Sigmoid activation function to obtain the output of SAM, which is finally

multiplied with the original map to revert to the C\*H\*W size. The CBAM can effectively focus on the local information region and improve the feature extraction capability.

### 3.2. Multi-scale Feature Fusion

#### 3.2.1. BiFPN Network

In the feature fusion part, the original network employs a top-down Feature Pyramid Network (FPN) structure to achieve feature fusion of shallow location information and deep semantic information. However, due to the unidirectional information flow, a bottom-up Path Aggregation Network (PANet) pathway is added to reduce information loss. Nevertheless, this structure lacks direct connections between two nodes at the same level, which limits the degree of multi-scale feature fusion that can be achieved. In this study, BiFPN structure is introduced without adding too much computational cost [32], as shown in **Error! Reference source not found.**. The importance of different input streams is determined by learnable weighting factors, while the additional cross-layer connection channel is added between two nodes to enable higher-level multi-scale feature fusion.

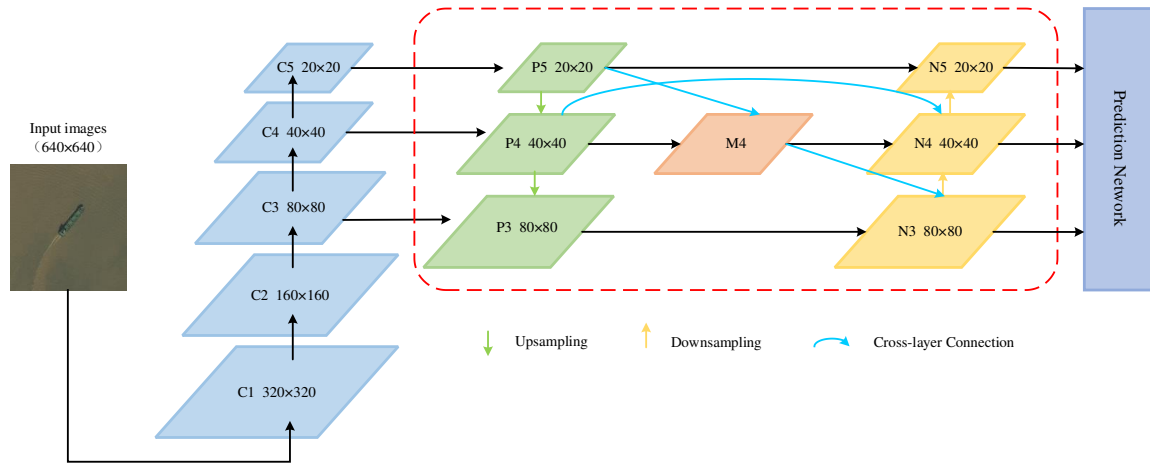


Figure 5. BiFPN feature fusion network.

The feature map C4 is chosen as an example to illustrate the feature fusion process based on the incorporation of cross-scale connection and contextual information weighting operations.

Firstly, each input stream requires a weighting factor to be assigned, and the weights are obtained by means of network self-learning, using the fast normalized fusion method [20] to constrain the size of each weight, which is calculated as shown:

$$O = \sum_i \frac{\omega_i \cdot X_i}{\varepsilon + \sum_j \omega_j} \quad (5)$$

where: O is the output feature;  $\omega_{i,j}$  is the learning weight coefficient of different layers;  $X_{i,j}$  is the input feature;  $\varepsilon \leq 0.001$ .

Secondly, it can be observed that the input streams of feature map N4 include P4, M4, and N3, where M4 is the image map of feature maps P4 and P5, and its calculation process is shown in Equation (6):

$$N_4 = \text{Conv} \left( \frac{\omega'_1 P_4^{in} + \omega'_2 P_4^{td} + \omega'_3 \cdot \text{Resize}(P_3^{out})}{\omega_1 + \omega_2 + \omega_3 + \varepsilon} \right) \quad (6)$$

where: Conv is the convolution operation;  $\omega_i$ ,  $\omega'_i$  are the different layer learning weights;  $P_i^{in}$ ,  $P_i^{out}$  are the input and output of layer i respectively;  $P_i^{td}$  is the output of the middle node of the feature



map in layer  $i$ ; Resize refers to maintaining the consistency of the input feature map size by using upsampling or downsampling operations;  $\varepsilon$  is usually set to 0.0001.

### 3.2.2. GSConv Structure

As the network deepens, the feature map gradually transmits spatial information to the channels during the backbone network feature extraction process. However, spatial compression and channel expansion operations lead to partial loss of semantic information. To address this issue, this paper uses a lightweight GSConv structure to replace the conventional Conv. This can alleviate the resistance of the input stream and reduce the complexity of the model, while greatly preserving each inter-channel hidden connections and spatial information without any compression processing. The structure is shown in **Error! Reference source not found..**

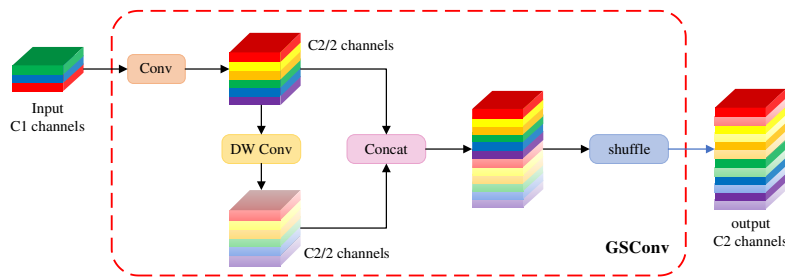


Figure 6. GSConv structure.

### 3.3. Loss Function

As an essential component of the object detection loss function, a well-designed localization loss function can significantly improve the performance of object detection models. Considering the differences in image quality in real scenes, it is inevitable that there will be low-quality examples in the training data. Geometric factors (such as distance and aspect ratio) will exacerbate the penalty for low-quality examples, thus reducing the generalization performance of the model. Based on the idea of dynamic non-monotonic focusing mechanism, Wise-IoU loss function is constructed to calculate the localization loss, which has three versions, namely W-IoUv1, W-IoUv2 and W-IoUv3 [22]. In this paper, utilizing the W-IoUv3 outlier degree to re-evaluate the quality of anchor boxes based on W-IoUv1, and assigning different gradient gains to samples of different qualities. **Error! Reference source not found.** illustrates a schematic diagram of the calculation of each parameter in the Wise-IoU loss.

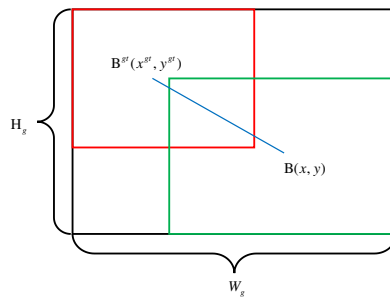


Figure 7. Schematic diagram of the calculation of each parameter in Wise-IoU loss.

The W-IoU<sub>v1</sub> loss function is defined as:

$$L_{W-IoU_{v1}} = R_{W-IoU} L_{IoU} \quad (7)$$

Where:

$$R_{W-IoU} = \exp \left( \frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*} \right)$$

$$L_{IoU} = \frac{B \cap B^{gt}}{B \cup B^{gt}}$$

where:  $(x, y)$  and  $(x_{gt}, y_{gt})$  are the coordinate values of the center points of the prediction and ground truth box;  $W_g, H_g$  are the width and height of the minimum bounding box between the prediction and ground truth box;  $B, B_{gt}$  are the areas of the prediction and ground truth box.

W-IoU<sub>v3</sub> introduces the outlier degree to describe the quality of the anchor box on based on W-IoU<sub>v1</sub>, it is defined as:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \quad (8)$$

A smaller outlier degree indicates a higher quality anchor box and is given a smaller gradient gain in order to concentrate the localization loss on regular anchor boxes; In addition, a larger outlier degree assigned a smaller gradient gain can effectively prevent harmful gradients from low quality examples.

Thus, the non-monotonic focusing mechanism factor  $r$  constructed using the outlier degree is applied to W-IoU<sub>v1</sub>, the definition of W-IoU<sub>v3</sub> is as follows:

$$L_{W-IoU_{v3}} = r L_{W-IoU_{v1}} \quad (9)$$

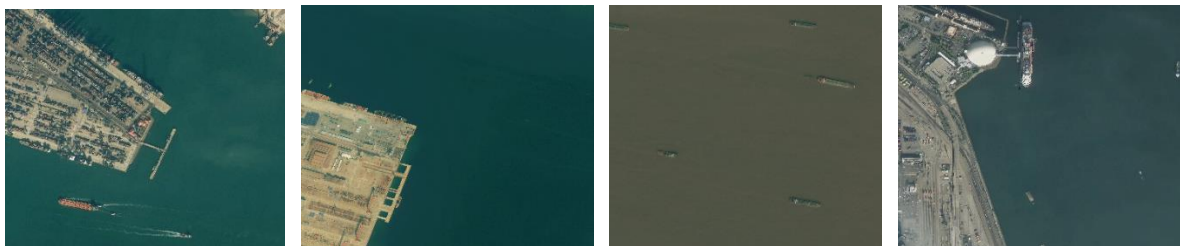
Where:

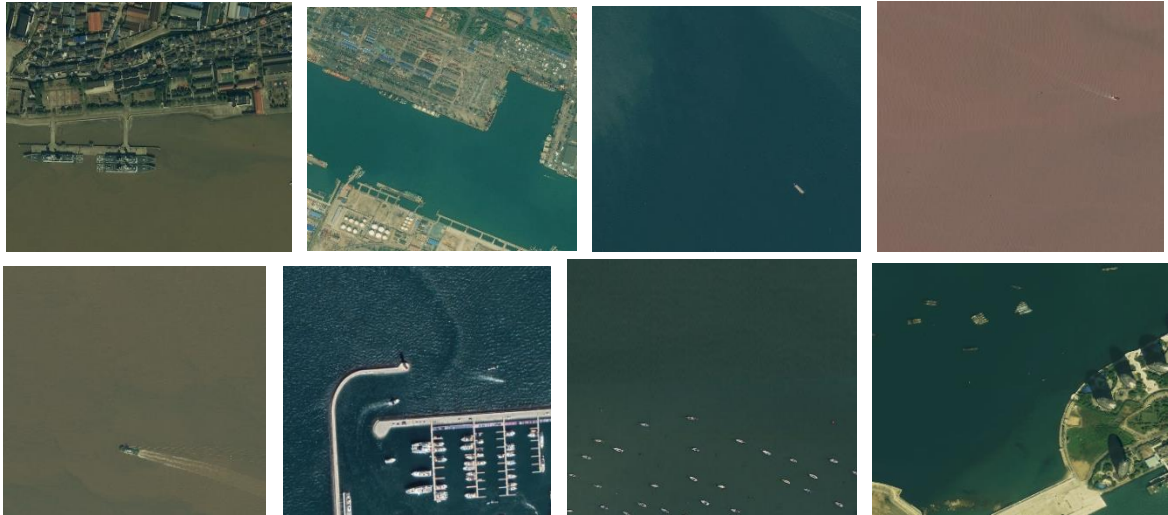
$$r = \frac{\beta}{\delta a^{\beta-\delta}}$$

## 4. Experimental Results and Analysis

### 4.1. Experimental Dataset

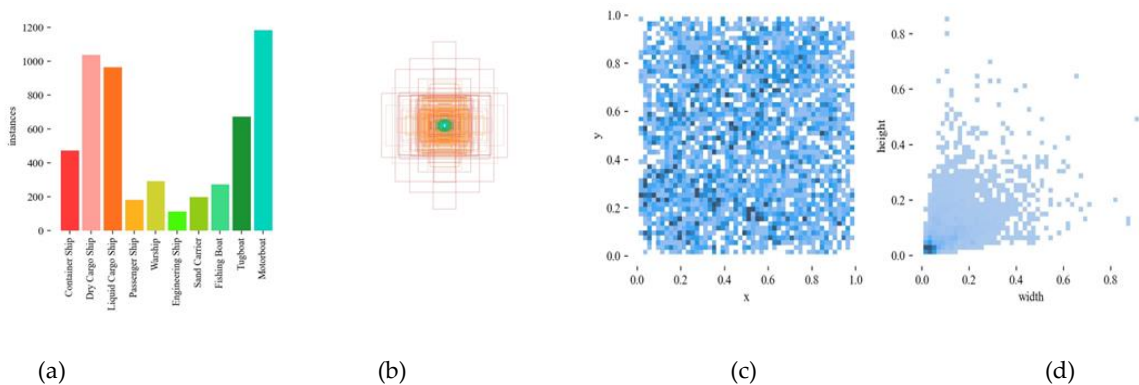
The dataset for this experiment is chosen from the Fine-grained Object Recognition in High-resolution Remote Sensing Imagery (FAIRIM) [33] created by the Aero-space Information Research Institute of the Chinese Academy of Sciences. The dataset covers various scenes such as nearshore ports and offshore areas. We have selected 2235 images and manually annotated to classify them into 10 common types of ships (Container Ship (CS), Dry Cargo Ship (DCS), Liquid Cargo Ship (LCS), Passenger Ship (PS), Warship (WS), Engineering Ship (ES), Sand Carrier (SC), Fishing Boat (FB), Tug-boat (TB), Motorboat (MB)). Some of example images in the FAIRIM dataset are shown in **Error! Reference source not found.**





**Figure 8.** Example images of the FAIRIM dataset.

The visualization results of the dataset are presented in **Error! Reference source not found..** Figure 9 (a) shows the types of vessels and the number of each category labelled in this dataset; Figure 9 (b) can see intuitively the distribution of anchor boxes for the data labels; Figure 9 (c) reveals the relative position of the detection targets compared to the whole image coordinate system. Figure 9 (d) is the normalized target size map, which means the size of the detection target in relation to the whole image, it can be seen from the figure that the target size distribution is relatively concentrated and mostly small.



**Figure 9.** Statistical results of the dataset: (a) bar chart of the number of targets in each class; (b) anchor box distribution map; (c) normalized target location map; (d) normalized target size map.

4.2. Experimental Platform and Parameters Setting

The experimental platform configuration used in this study is shown in **Error! Reference source not found..**

**Table 1.** Experimental environment configuration.

Parameter	Configuration
Operating Environment	Windows 11
GPU	GeForce RTX 3050
Programming Language	Python 3.7
Programming Platform	Pycharm
Deep Learning Framework	Pytorch 1.13.0
CUDA	11.0

CuDNN	8.0
-------	-----

The network model training was conducted under the experimental environment described above, and the training parameters are shown in **Error! Reference source not found..**

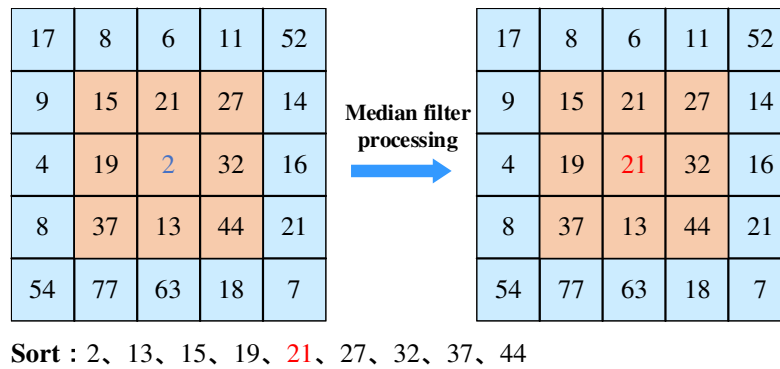
**Table 2.** Experimental parameters setting.

Parameter	Number
img-size	640×640
batch-size	8
epochs	300
learning rate	0.01
momentum	0.937
weight-decay	0.0005

#### 4.3. Image Preprocessing

To eliminate the interference of water ripples, waves, wakes and salt and pepper noise around the ship and highlight the information of ship features, this paper combines median and bilateral filter methods for noise reduction of the images.

Median Filter (MF) involves replacing the pixel value of a point in an image with the median value of the values of all points in its neighborhood. This makes the surrounding pixel values closer to the true values, thereby eliminating isolated noise points [34], An example of this is shown in **Error! Reference source not found..**



**Figure 10.** Example of median filter processing.

Bilateral Filter (BF), as a non-linear filter that takes into account the influence weight of the Euclidean distance and the grey scale interpolation when calculating the new value of a certain pixel point [35]. It can effectively remove noise and preserve the edge features of ships better. The definition is as follows:

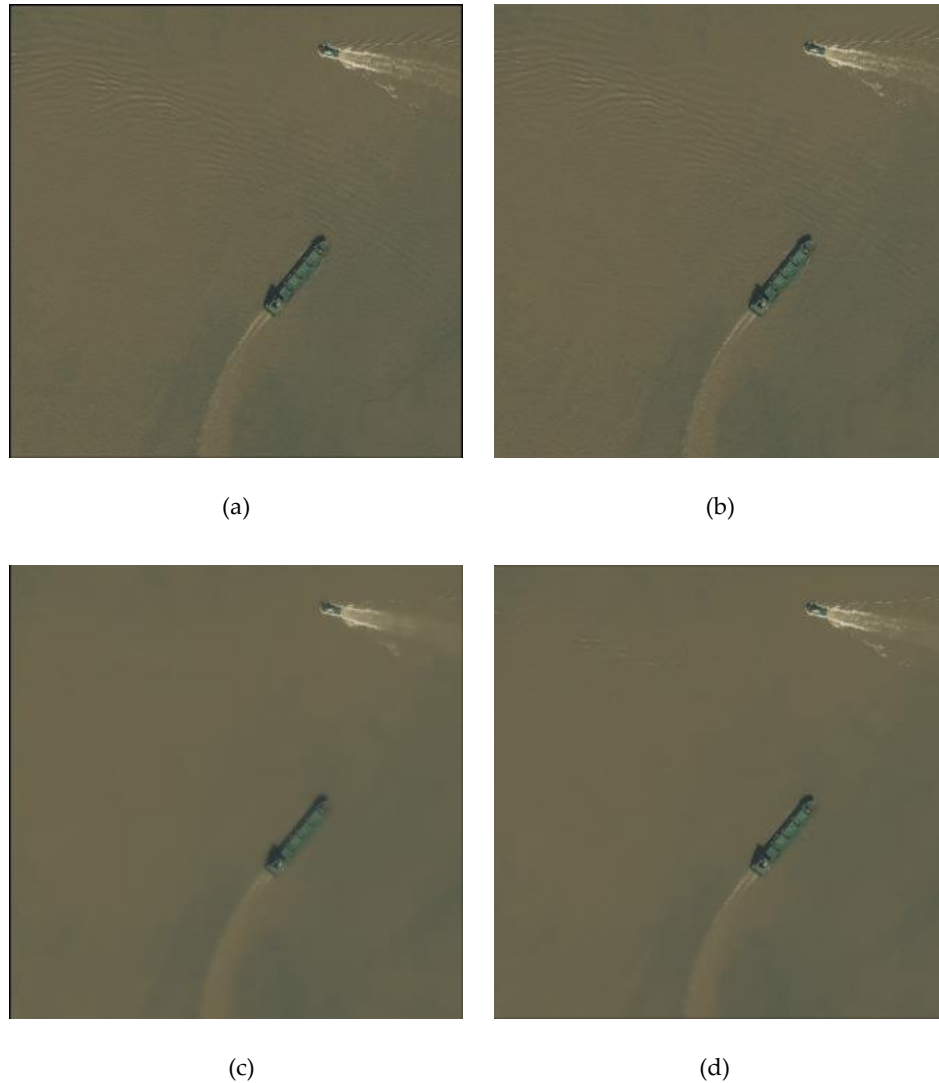
$$BF[I]_p = \frac{1}{W_p} \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(\|I(p) - I(q)\|) I(q) \quad (10)$$

Where:

$$W_p = \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(\|I(p) - I(q)\|)$$

where:  $BF[I]_p$  is the image after bilateral processing;  $W_p$  is the normalization factor;  $S$  is the neighborhood of a pixel point;  $G_{\sigma_s}(\|p - q\|)$  is the spatial weight;  $\sigma_s$  is the standard deviation of the spatial domain;  $p, q$  are any two pixel points in the image;  $G_{\sigma_r}(\|I(p) - I(q)\|)$  is the pixel range weight;  $\sigma_r$  is the standard deviation of the range;  $I(p)$  is the input image;  $I(q)$  is the output image.

Through comparison between the median filter, bilateral filter and their combination, it can be found that the combination of median filter and bilateral filter has a better noise reduction effect than using them separately, so this paper adopts the median + bilateral filter method for image noise reduction of the dataset, and its noise reduction effect is shown in **Error! Reference source not found..**



**Figure 11.** Median filter and bilateral filter for noise reduction: (a) original image; (b) median filter for noise reduction; (c) bilateral filter for noise reduction; (d) median + bilateral filter for noise re-duction.

#### 4.4. Evaluation Metrics

In this study, Average Precision (AP), mean Average Precision (mAP) and Frame Per Second (FPS) are used as evaluation metrics to assess the performance of the algorithm.

The AP is a measure of the quality of detection results for a certain category, usually related to Recall (R) and Precision (P); The mAP is a measure of detection results for multiple categories, and its calculation is related to the AP; The specific Equation for calculating AP and mAP are as follows.

Recall (R) is calculated by the Equation:

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

where:  $N_{TP}$  is the number of positive samples that are correctly classified;  $N_{FN}$  is the number of samples that are positive but are misclassified as negative.

Precision (P) is calculated by the Equation:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

where:  $N_{FP}$  is the number of samples that are negative but are misclassified as positive.

The AP is the area enclosed by the P-R curve, with Precision and Recall as the vertical and horizontal coordinates, which is calculated as:

$$AP = \int_0^1 P(r)dr \tag{11}$$

The AP of all samples were then averaged to obtain the mAP, which is calculated as:

$$mAP = \frac{\sum AP}{N_{all\_classes}} \tag{12}$$

where:  $N_{all\_classes}$  is the total number of samples. mAP\_0.5 and mAP\_0.5:0.95 are two representations of mean Average Precision. mAP\_0.5 sets the IoU threshold to 0.5 and then averages the AP for all classes; mAP\_0.5:0.95 sets different thresholds ranging from 0.5 to 0.95 with a step size of 0.5, and then averages the AP for all classes.

The FPS stands for frames per second, is the main indicator for real-time performance evaluation of a model, which is calculated as:

$$FPS = \frac{1000}{t_p + t_i + t_{NMS}} \tag{13}$$

where:  $t_p$  is the image preprocessing time;  $t_i$  is the model inference time;  $t_{NMS}$  is the Non-Maximum Suppression time.

4.5. Ablation Experiment

In order to verify the detection performance of the improved method, ablation experiments were conducted on each improved component, and compared experimentally with the detection performance of the original model. The roles of each improvement strategy are as follows: Adding the CBAM effectively suppresses useless information and focuses more on the target information; Replacing the BiFPN network enables feature fusion at different levels; Replacing the conventional Conv with the GSConv structure reduces the number of model parameters and shortens the training time; The target's height and width in pixel values and categorize its dataset into Large-size (256 pixels × 256 pixels), Medium-size (128 pixels × 128 pixels), and Small-size (64 pixels × 64 pixels) [36]. The results of the ablation experiment are shown in **Error! Reference source not found.**, with the best results highlighted in bold.

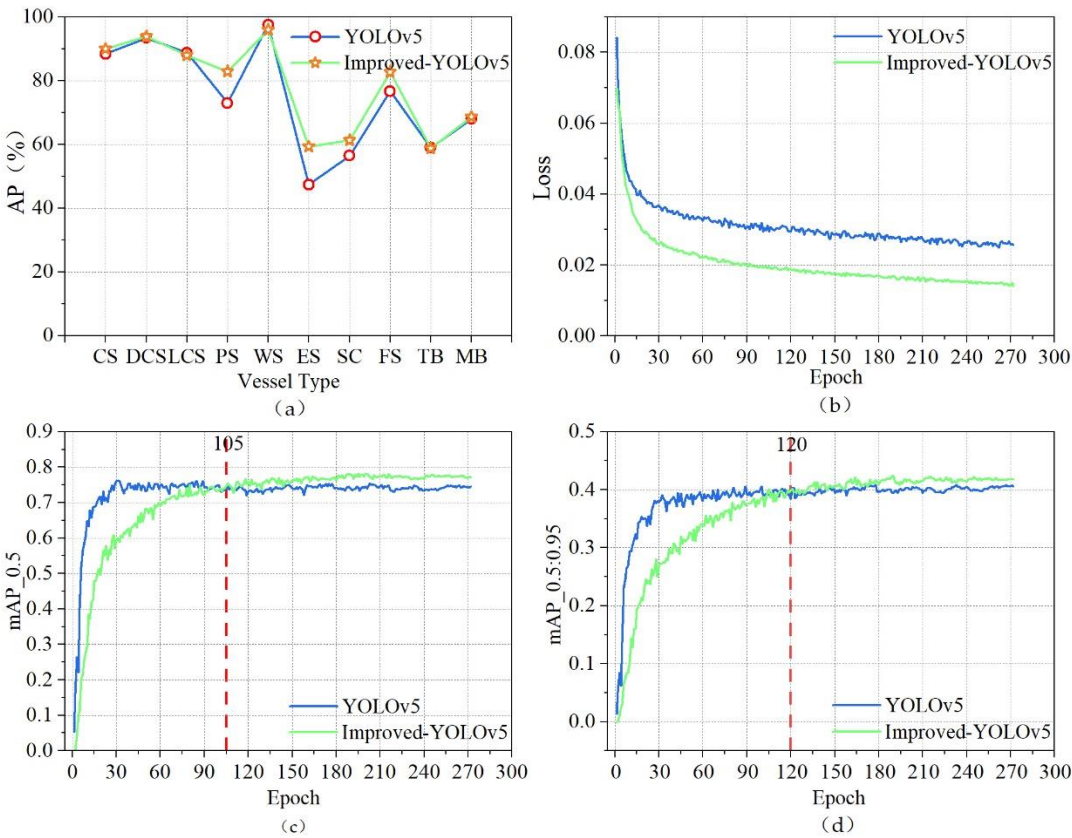
Table 3. Detailed comparison of the results of the ablation experiment.

Improvement Strategy			AP (%)										mAP_0.5 (%)	FPS (f/s)
			Large-size				Medium-size			Small-size				
CBA M	BiFPN+GSCo nv	W-Io U	CS	DC S	LC S	PS	W S	ES	SC	FB	TB	M B		
×	×	×	88.4	93.4	88.7	73	97.6	47.4	56.5	76.7	59	68	74.9	69
✓	✓	×	91.2	93.9	90.4	81.9	95.4	55.2	55.9	84.9	53.5	66.5	76.9	76
✓	×	✓	89.3	92.9	87.6	77.8	97.4	56.2	58.1	79	47.5	63.8	75	76



×	√	√	89. 2	<b>94.7</b>	88. 9	78	<b>97.</b> 7	<b>61.</b> 7	47. 4	81. 7	54	64. 7	75.8	<b>76</b>
√	√	√	89. 9	93.9	87. 9	<b>82.</b> 9	96	59. 3	<b>61.</b> 3	82. 7	58. 8	<b>68.</b> 6	<b>78.1</b>	75

As can be seen from Table 3, in the ablation experiments, the improvement strategies proposed in this paper have achieved certain improvements in the detection performance indicators compared to the original model. Adding CBAM and BiFPN+GSConv increased mAP\_0.5 by 2%; Adding CBAM and W-IoU increased mAP\_0.5 by 0.1%; Adding BiFPN+GSConv and W-IoU increased mAP\_0.5 by 0.9%; After adding all the improvement strategies, the overall mAP\_0.5 improved by 3.2%. Although the detection speed is slightly reduced compared to the combination of the other two strategies, it does not affect the real-time target detection requirements. Ultimately, the improvement scheme was determined to include the CBAM, BiFPN network, lightweight GSConv structure and the W-IoU loss function. The experimental results before and after its improvement with the original model YOLOv5 are shown in **Error! Reference source not found..**

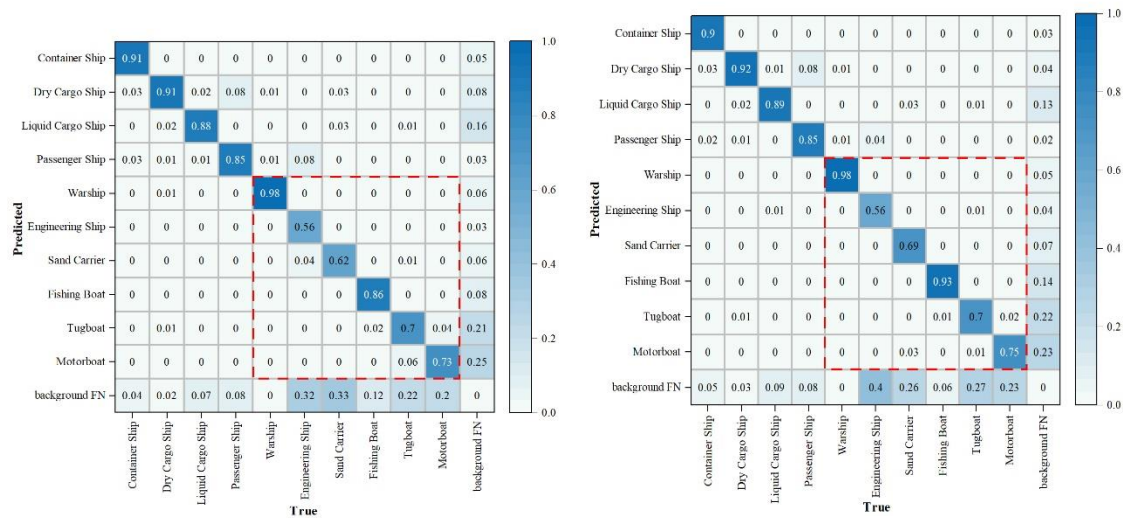


**Figure 12.** Comparison of experimental results before and after algorithm improvement: (a) AP by vessel type; (b) overall loss of model; (c)mAP\_0.5; (d)mAP\_0.5:0.95.

As shown in Figure 12, Improved-YOLOv5 has a more significant improvement in average precision for small and medium ships, and the overall loss of the model is less compared to the original YOLOv5. In terms of mean Average Precision, mAP\_0.5 and mAP\_0.5:0.95 first surpassed the original YOLOv5 at about 105 and 120 epochs respectively, and have maintained the lead ever since. In conclusion, after 300 epochs of training, the Improved-YOLOv5 algorithm has achieved better evaluation metrics than the original YOLOv5.

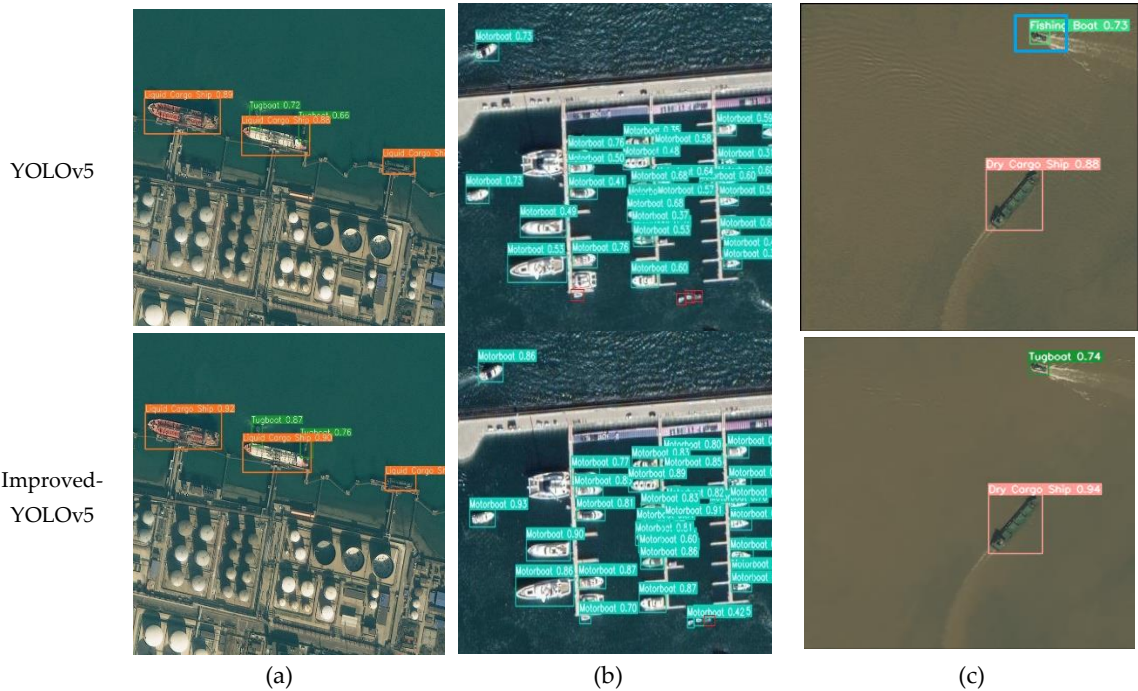
A confusion matrix was utilized to evaluate the accuracy of the model’s results. Each column of the confusion matrix represents the predicted proportions of each category, while each row represents the true proportions of the respective category in the data [37]. The confusion matrix for the two algorithms before and after the improvement is shown in **Error! Reference source not found..**

As can be seen from Figure 13, most of the targets were correctly predicted. The improved-model has slightly better detection performance than the original model on small and medium ship targets.



**Figure 13.** The confusion matrix of the before and after model improvement: (a) YOLOv5; (b) Improved-YOLOv5.

The comparison of the detection performance before and after the algorithm improvement is shown in **Error! Reference source not found..** By comparing the pictures, it can be found that the improved algorithm can effectively improve the detection effect, avoid the situation of missed detection and false detection, and at the same time improve the detection ability of small targets. The following image is displayed enlarged based on the original image.



**Figure 14.** Comparison of the detection performance before and after the algorithm improvement: (a) multi-scale; (b) small targets and tight arrangement; (d) disturbed by water wave background (The red box indicates missed detections, while the blue box indicates false detections).

4.6. Performance Comparison of Various Target Detection Algorithms

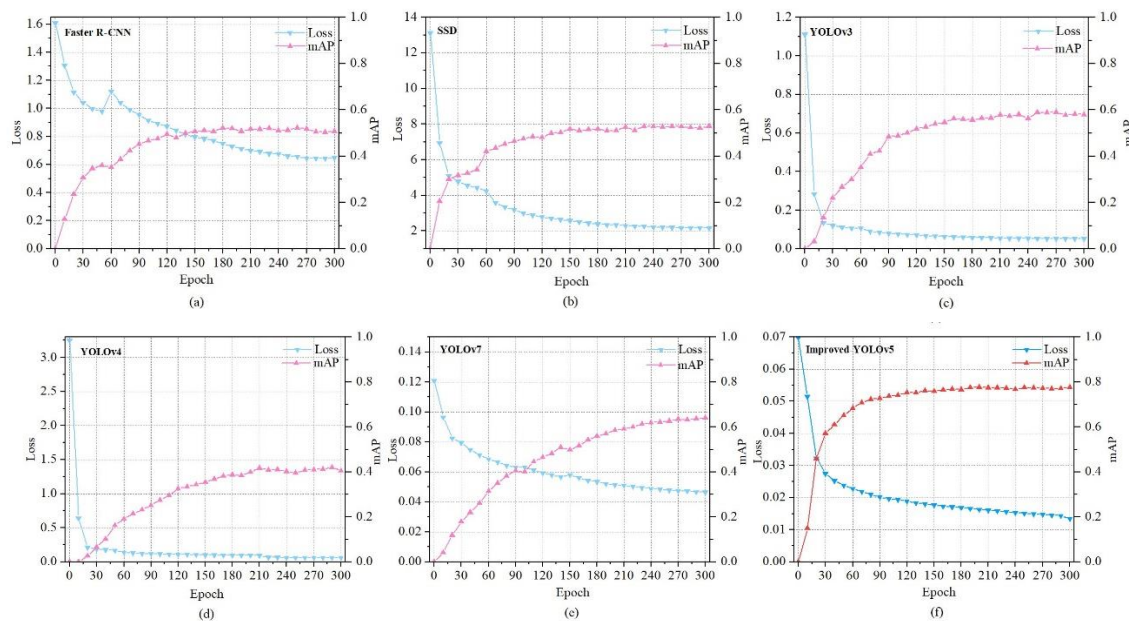
In order to further validate the effectiveness of the improved algorithm, this study compared it with several mainstream target detection algorithms, including Faster R-CNN, SSD, YOLOv3,

YOLOv4 and YOLOv7. The results of the comparison experiments are shown in **Error! Reference source not found.**, with the best results highlighted in bold.

**Table 4.** Comparison of detection results between different algorithms.

Modeling Algorithm	AP (%)										mAP (%)	FPS (f/s)
	Large-size			Medium-size				Small-size				
	CS	DCS	LCS	PS	WS	ES	SC	FB	TB	MB		
Faster R-CNN	95.7	76.3	81.6	72	94.6	48.2	16.4	25.9	10.1	16.9	53.8	7
SSD	92.3	73.9	82.3	77.3	88.3	37.1	16.3	73.8	19.5	19.3	57.9	45
YOLOv3	93.9	72.7	84.6	85	92	26.2	10.2	55.7	38.7	50.8	60.9	33
YOLOv4	78.2	69.5	79.1	72.3	77.4	0	0	0	12.9	42.5	43.2	41
YOLOv7	84.3	89.9	73.9	76.8	96.4	34	26.5	79.1	27	54.6	64.3	35
Improved-YOLOv5	89.9	93.9	87.9	82.9	96	59.3	61.3	82.7	58.8	68.6	78.1	75

From the experimental results, it can be observed that the Improved-YOLOv5 proposed in this study achieves the best detection performance on the FAIRIM dataset, with the mAP of 78.1%. Improved-YOLOv5 has increased 24.3%, 20.2%, 17.2%, 34.9% and 13.8% in the mAP compared to Faster R-CNN, SSD, YOLOv3, YOLOv4 and YOLOv7 respectively. In addition, it has an excellent performance in detecting small and medium ships. Furthermore, the detection rate is significantly improved compared to other model algorithms, which better meet the requirements for real-time target detection. The experimental comparison results of training loss and the mAP of Improved-YOLOv5 and various target detection algorithms are shown in **Error! Reference source not found.**



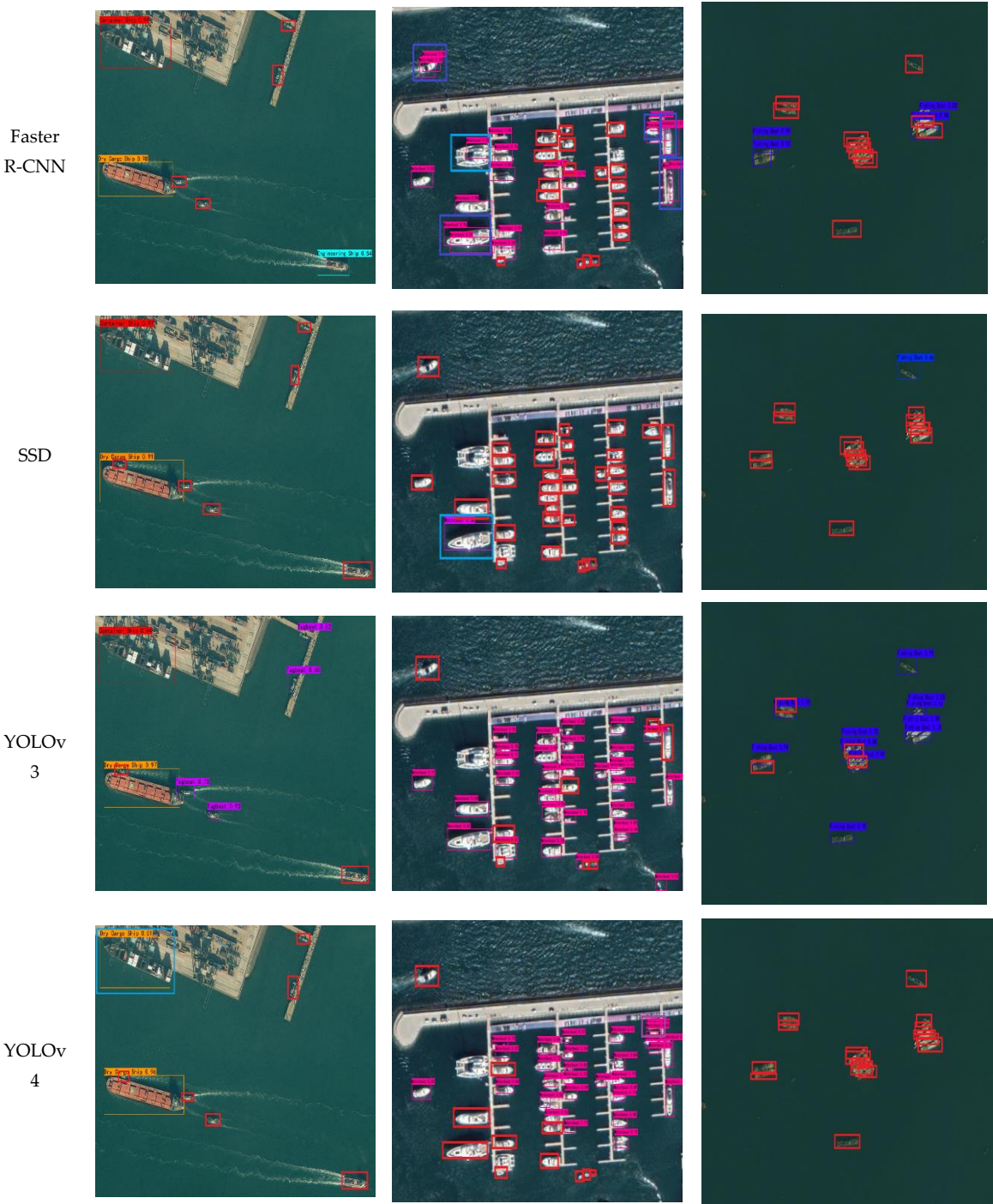
**Figure 15.** Comparison results of training loss and the mAP of Improved-YOLOv5 and various target detection algorithms: (a) Faster R-CNN; (b) SSD; (c) YOLOv3; (d) YOLOv4; (e) YOLOv7; (f) Improved-YOLOv5.

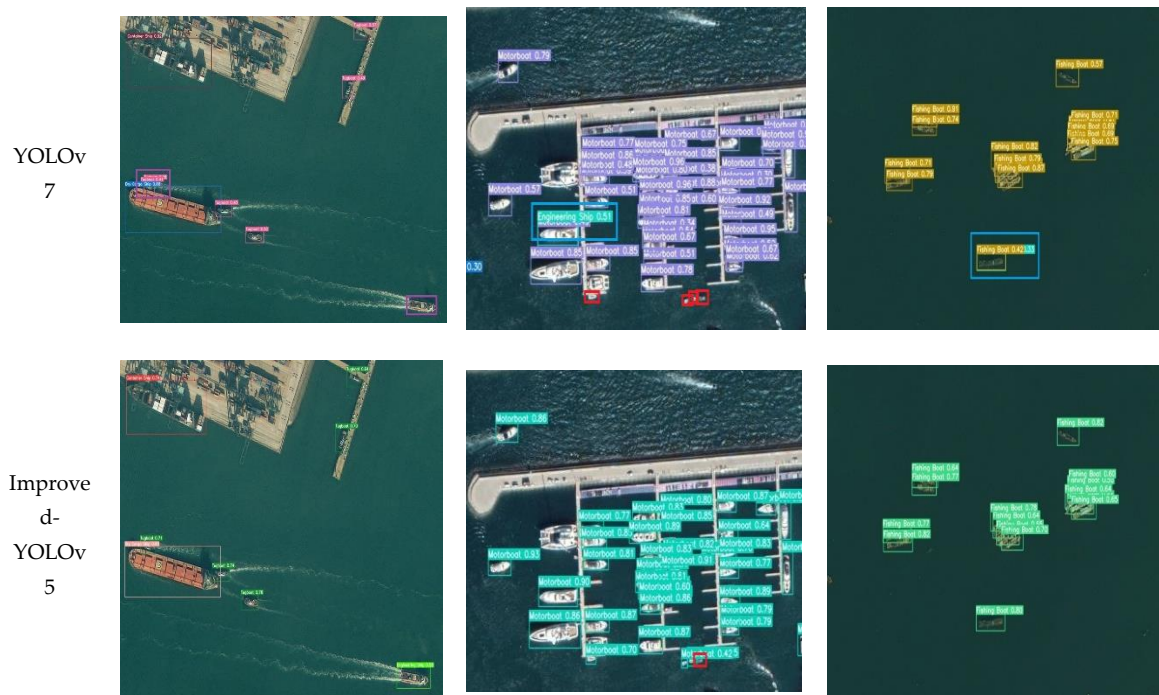
#### 4.7. Comparison of detection results

Due to the special characteristics of the distance and angle of remote sensing satellite imaging, the size of ship targets in the entire image is relatively small, resulting in limited feature information in the network extraction process, which leads to unsatisfactory detection performance of the current various target detection algorithms. In order to effectively verify the detection performance of our algorithm and several commonly used algorithms in complex scenes, we compare the proposed



method with 5 typical methods (Faster R-CNN, SSD, YOLOv3, YOLOv4, and YOLOv7). **Error! Reference source not found.** intuitively shows the ship detection results in different maritime scenes. Complex scenes include situations such as multi-scale, a large number of small targets and tight arrangement. Compared to the others, our proposed improved method can efficiently and accurately identify ship types in different scenes captured by remote sensing satellites, especially in the detection ability of small objects. The following image is displayed enlarged based on the original image.





**Figure 16.** Comparison of detection performance between Improved-YOLOv5 and various algorithms ((Faster R-CNN, SSD, YOLOv3, YOLOv4, and YOLOv7), the red box indicates missed detections, the blue box indicates false detections, and the purple box indicates overlapping anchor boxes).

The comparison in Figure 16 shows that Improved-YOLOv5 outperforms other algorithms in terms of both missed and false detections, and at the same time has a high detection accuracy. Therefore, it can meet the requirements of accurate and real-time detection of ships on the sea surface from remote sensing images.

## 5. Conclusions

Ship target detection in remote sensing images faces challenges such as complex scenes, tight arrangement and mostly small targets, which can result in poor overall detection performance and even missed or false detection.

Therefore, we propose the Improved-YOLOv5 algorithm. The contribution of this study is threefold. Firstly, adding the CBAM in the backbone network to adaptively optimize the features of the targets in both channel and spatial dimensions, enabling it to focus more on the relevant feature regions. Secondly, the BiFPN network is used in the feature fusion process to achieve higher-level multi-scale feature fusion by cross-layer connecting input and output nodes of the same level; At the same time, the lightweight GSConv structure replaces the conventional Conv, reducing the model's parameter and computational complexity. Finally, to address sample image quality differences and reduce the harmful gradients of low-quality examples, the Wise-IoU loss function is applied at the output for calculating the localization loss and assigning reasonable gradient gains. After experimental comparison, the improved algorithm in this paper has improved the mAP by 3.2% and accelerated the FPS by 8.7% compared to the original YOLOv5s, which can effectively improve the detection performance of small targets and reduce the missed detection rate of densely distributed small targets.

In our future research, the next step will be to continue optimizing the network architecture to improve ship detection performance in complex remote sensing images. This also involves reducing network inference complexity and optimizing anchor box orientation and size to make it better portable for deployment on embedded platforms.

**Author Contributions:** Conceptualization, J.J., L.L., Y.Z., K. X.; methodology, J.J., L.L.; software, L.L.; validation, L.L.; formal analysis, J.J., L.L.; investigation, J.J., Y.Z.; resources, J.J.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, J.J., J.Y.; visualization, L.L., K. X., J.Y.; supervision, J.J.; project administration, J.J.; funding acquisition, J.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant No. 42261144671, 42030602, 41725018).

**Data Availability Statement:** The data used to support the findings of this study are available from the corresponding author upon request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Wang, B.; Dong, L.L.; Zhao, M.; Wu, H.D.; Ji, Y.Y.; Xu, W.H. An infrared maritime target detection algorithm applicable to heavy sea fog. *Infrared Phys. Technol.* **2015**, *71*, 56–62. [CrossRef]
2. Zhao, E.Z.; Dong, L.L.; Dai, H. Infrared Maritime Small Target Detection Based on Multidirectional Uniformity and Sparse-Weight Similarity. *Remote Sens.* **2022**, *14*, 5492. [CrossRef]
3. Yang, P.; Dong, L.; Xu, H.; Dai, H.; Xu, W. Robust Infrared Maritime Target Detection via Anti-Jitter Spatial-Temporal Trajectory Consistency. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
4. Lang, H.T.; Wang, R.F.; Zheng, S.Y.; Wu, S.W.; Li, J.L. Ship Classification in SAR Imagery by Shallow CNN Pre-Trained on Task-Specific Dataset with Feature Refinement. *Remote Sens.* **2022**, *14*, 5986. [CrossRef]
5. Liu, P.F.; Wang, Q.; Zhang, H.; M, J.; Liu, Y.C. A Lightweight Object Detection Algorithm for Remote Sensing Images Based on Attention Mechanism and YOLOv5s. *Remote Sens.* **2023**, *15*, 2429. [CrossRef]
6. Nie, G.T.; Huang, H. A Survey of Object Detection in Optical Remote Sensing Images. *Acta Auto. Sin.* **2021**, *47*, 1749–1768.
7. Li, K.; Fan, Y. Research on Ship Image Recognition Based on Improved Convolution Neural Network. *Ship Sci. Tech.* **2021**, *43*, 187–189.
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV), Lake City, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
9. Girshick, R. Fast R-CNN. In Proceedings of the IEEE international Conference on Computer Vision (ICCV), Santiago, Chile, 07–13 December 2015; pp. 1440–1448. [CrossRef]
10. Ren, S.Q.; He K.M.; Girshick, R.; Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the IEEE transactions on pattern analysis and machine intelligence, 06 June 2016; pp. 1137–1149. [CrossRef]
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision, Springer, The Netherlands, 17 September 2016; pp. 21–37. [CrossRef]
12. Redmon, J.; Kumar, S.; Divvala, K.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
13. Zhang, P.P.; Xie, G.K.; Zhang, J.S. Gaussian Function Fusing Fully Convolutional Network and Region Proposal-Based Network for Ship Target Detection in SAR Images. *Int. J. Ante. Prop.* **2022**. [CrossRef]
14. Wen, G.Q.; Cao, P.; Wang, H.N.; Chen, H.L.; Liu, X.L.; Xu, J.H.; Zaiane, O. MS-SSD: multi-scale single shot detector for ship detection in remote sensing image. *Applied Int.* **2022**, *53*, 1586–1604 [CrossRef]
15. Chen, L.Q.; Shi, W.X.; Deng, D.X. Improved YOLOv3 Based on Attention Mechanism for Fast and Accurate Ship Detection in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 660. [CrossRef]
16. Huang, Z.X.; Jiang, X.N.; Wu, F.L.; Fu, Y.; Zhang, Y.; Fu, T.J.; Pei, J.Y. An Improved Method for Ship Target Detection Based on YOLOv4. *Applied Sci.* **2023**, *13*, 1302. [CrossRef]
17. Zhou, J.C.; Jiang, P.; Zou, A.; Chen, X.L.; Hu, W.W. Ship Target Detection Algorithm Based on Improved YOLOv5. *J. Mar. Sci. Eng.* **2021**, *9*, 908. [CrossRef]
18. Dong, Z.; Lin, B.J. Learning a robust CNN-based rotation insensitive model for ship detection in VHR remote sensing images. *Int. J. Remote Sens.* **2020**, *41*, 3614–3626. [CrossRef]
19. Woo, S.H.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. ECCV. In Proceedings of the European Conference on Computer Vision, New York, USA, 2018; pp. 3–19. [CrossRef]
20. Tan, M.; Pnag, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, 13–19 June 2020; pp.10778–10787. [CrossRef]
21. Liang, X.M.; Xiao, H.; Lightweight Strip Defect Real-time Detection Algorithm Based on SDD-YOLO. *China Meas. Test* **2023**, *18*.



22. Tong, Z.J.; Chen, Y.H.; Xu Z.W.; Yu R. Wise-IOU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. In Proceedings of the IEEE Transactions on Circuits and Systems for Video Technology. February 2023. [CrossRef]
23. Wang, Z.; Wu, L.; Li, T.; Shi, P.B. A Smoke Detection Model Based on Improved YOLOv5. *Math.* **2022**, *10*, 1190. [CrossRef]
24. Malta, A.; Mendes, M.; Farinha, T. Augmented Reality Maintenance Assistant Using YOLOv5. *Appl. Sci.* **2021**, *11*, 4758. [CrossRef]
25. Wang, C.Y.; Markliao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA, 14-19 June 2020; pp. 390-391 [CrossRef]
26. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 21-26 July 2017; pp. 2117-2125. [CrossRef]
27. Liu, S.; Qi, L.; Qin, H.F.; Shi, J.P.; Jia, J.Y. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 18-23 June 2018; pp. 8759-8768. [CrossRef]
28. Zhang, X.P.; Xu, Z.Y.; Qu, J.; Qiu, W.X.; Zhai, Z.Y. Maritime Ship Recognition Based on Improved YOLOv5 Deep Learning Algorithm. *J. Dalian Ocean Univ.* **2022**, *37*, 866-872.
29. Jiang, B.R.; Luo, R.X.; Mao, J.Y.; Xiao, T.; Jiang, Y.N. Acquisition of localization confidence for accurate objection. In Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 8-14 September 2018; pp. 784-799. [CrossRef]
30. Merugu, S.; Tiwari, A.; Sharma, S.K. Spatial-Spectral Image Classification with Edge Preserving Method. *J. Indian Soc. Remote Sens.* **2021**, *49*, 703-711. [CrossRef]
31. Fu, G.D.; Huang, J.; Yang, T.; Zheng, S.Y.; Improved Lightweight Attention Model Based on CBAM. *Comput. Eng. Appl.* **2021**, *57*, 150-156.
32. Zhao, W.Q.; Kang, Y.J.; Zhao, Z.B.; Zhai, Y.J. A Remote Sensing Image Object Detection Algorithm with Improved YOLOv5s. *CAAI Trans. Int. Sys.* **2023**, *18*, 86-95.
33. Sun, X.; Wang, P.J.; Yan, Z.Y.; Xu, F.; Wang, R.P.; Diao, W.h.; Chen, J.; Li, J.H.; Feng, Y.C.; Xu, T.; Weinmann, M.; Hinz, S.; Wang, C.; Fu, K. ISPRS Journal of Photogrammetry and Remote Sensing. *ISPRS J. Phot. Remote Sens.* **2022**, *184*, 116-130. [CrossRef]
34. Kumar, S.; Raja, R.; Mahmood, M. R.; Choudhary, S. A Hybrid Method for the Removal of RVIN Using Self Organizing Migration with Adaptive Dual Threshold Median Filter. *Sens. Imaging* **2023**, *24*. [CrossRef]
35. Wu, L.S.; Fang, L.Y.; Yue, J.; Zhang, B.; Ghamisi, P.; He, M. Deep Bilateral Filtering Network for Point-Supervised Semantic Segmentation in Remote Sensing Images. *IEEE Tran. Image Proc.* **2022**, *31*, 7419-7434. [CrossRef]
36. Gong, H.; Mu, T.K.; Li, Q.X.; Dai, H.S.; Li, C.L.; He, Z.P.; Wang, W.J.; Han, F.; Tuniyani, A.; Li, H.Y.; Lang, X.C.; Li, Z.Y.; Wang, B. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
37. Liu, K.Y.; Sun, Q.; Sun, D.M.; Peng, L.; Yang, M.D.; Wang, N.Z. Underwater Target Detection Based on Improved YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 667. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.